

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2002年11月11日

出 願 番 号

Application Number:

特願2002-326598

[ ST.10/C ]:

[ JP2002-326598 ]

出 願 人

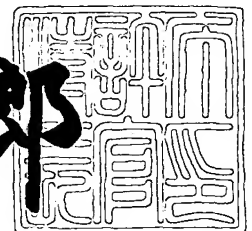
Applicant(s):

株式会社日立製作所

2003年 6月13日

特許庁長官  
Commissioner,  
Japan Patent Office

太田信一郎



出証番号 出証特2003-3046492

【書類名】 特許願

【整理番号】 HI020566

【提出日】 平成14年11月11日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 3/06

【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 RAIDシステム事業部内

【氏名】 高田 豊

【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 RAIDシステム事業部内

【氏名】 小林 直孝

【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 RAIDシステム事業部内

【氏名】 小笠原 裕

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100071283

【弁理士】

【氏名又は名称】 一色 健輔

【選任した代理人】

【識別番号】 100084906

【弁理士】

【氏名又は名称】 原島 典孝

## 【選任した代理人】

【識別番号】 100098523

【弁理士】

【氏名又は名称】 黒川 恵

## 【選任した代理人】

【識別番号】 100112748

【弁理士】

【氏名又は名称】 吉田 浩二

## 【選任した代理人】

【識別番号】 100110009

【弁理士】

【氏名又は名称】 青木 康

## 【手数料の表示】

【予納台帳番号】 011785

【納付金額】 21,000円

## 【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 ディスク制御装置およびディスク制御装置の制御方法

【特許請求の範囲】

【請求項 1】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段と、

前記ネットワーク制御部が複数のアドレスが設定されている 1 の前記コマンドをディスク制御部に送信する手段と、

前記ディスク制御部が前記コマンドを受信してこのコマンドに設定されている前記各アドレスに対応するデータ入出力をディスクドライブに対して行う手段と、

を備えることを特徴とするディスク制御装置。

【請求項 2】 請求項 1 に記載のディスク制御装置において、

前記ネットワーク制御部ではファイルシステムが動作し、

前記データ入出力要求は、ディスクドライブに入出力されるデータをファイル名により指定するものであり、

前記ネットワーク制御部は、前記データ入出力要求に設定されているファイル名に対応するデータのディスクドライブ上の記憶位置に対応するアドレスを生成し、このアドレスを前記コマンドに設定する手段を備えること、

を特徴とするディスク制御装置。

【請求項 3】 請求項 1 に記載のディスク制御装置において、

前記アドレスは前記ディスクドライブの記憶領域に区画設定された論理的な記憶領域におけるデータの記憶位置を指定する論理的なアドレスであること、

を特徴とするディスク制御装置。

【請求項 4】 請求項 1 に記載のディスク制御装置において、前記内部バス

は P C I バスであること、

を特徴とするディスク制御装置。

【請求項 5】 請求項 1 に記載のディスク制御装置において、前記ネットワーク制御部はネットワークプロトコルに従って前記外部装置と通信する手段を備えること、

を特徴とするディスク制御装置。

【請求項 6】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段を備え、

前記回路基板に前記ネットワーク制御部および前記ディスク制御部が共通してアクセス可能なメモリを有し、

前記ネットワーク制御部および前記ディスク制御部が設定されたタイミングで前記メモリに自身の動作状態を示す動作状態情報を更新する手段を備え、

前記動作状態情報に基づいて前記ネットワーク制御部もしくは前記ディスク制御部に障害が発生したことを検知する手段を備えること、

を特徴とするディスク制御装置。

【請求項 7】 請求項 6 に記載のディスク制御装置において、

前記ネットワーク制御部は、前記コマンドを前記ディスク制御部に送信するに際し前記コマンドの送信先となる前記ディスク制御部の動作状態を前記動作状態情報から取得し、取得した動作状態に応じて前記コマンドを前記ディスク制御部に送信するかどうかを決定すること、

を特徴とするディスク制御装置。

【請求項 8】 請求項 6 に記載のディスク制御装置において、

前記ネットワーク制御部は、前記ディスク制御部に送信したコマンドについての受信通知を取得できない場合に前記動作状態情報に基づいて送信先の前記ディ

スク制御部の動作状態を調査し、その調査の結果に応じてそのコマンドを再び前記ディスク制御部に送信するかどうかを判断すること、

を特徴とするディスク制御装置。

【請求項 9】 請求項 6 に記載のディスク制御装置において、

前記ネットワーク制御部は、前記ディスク制御部に送信したコマンドについての受信通知を取得できない場合に前記動作状態情報に基づいて送信先の前記ディスク制御部の動作状態を調査し、前記ディスク制御装置が正常に動作していないと判断した場合には、他の前記ディスク制御部に前記コマンドを送信すること、

を特徴とするディスク制御装置。

【請求項 10】 請求項 6 に記載のディスク制御装置において、

前記障害を検知した場合にその旨を通知するユーザインタフェースを備えることを特徴とするディスク制御装置。

【請求項 11】 請求項 6 に記載のディスク制御装置において、

前記障害を検知した場合に、障害が発生している前記ネットワーク制御部もしくは前記ネットワーク制御部に、再起動を要求する信号を送信すること、

を特徴とするディスク制御装置。

【請求項 12】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段を備え、

前記ディスク制御部がバックアップ装置との接続手段を備え、

前記ネットワーク制御部が前記ディスクドライブに記憶しているデータについてのバックアップ要求を外部装置から受信して前記ディスク制御部にバックアップコマンドを送信する手段を備え、

前記ディスク制御部は前記バックアップコマンドを受信すると前記ディスクドライブのデータについてのバックアップ指示をバックアップ装置に送信する手

段を備えること、

を特徴とするディスク制御装置。

【請求項 1 3】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うディスク制御装置において、

互いに通信可能に接続された複数の前記回路基板を備え、

前記回路基板間でハートビートメッセージを交換することで 1 の前記回路基板に障害が生じたことを他の回路基板が検知する手段を備え、

前記回路基板が他の前記回路基板についての障害を検知した場合に、障害となっている前記回路基板が行っている処理をその回路基板とは異なる他の回路基板に代行させる手段を備えること、

を特徴とするディスク制御装置。

【請求項 1 4】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段を備えることを特徴とするディスク制御装置。

【請求項 1 5】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、前記ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うディスク制御装置を、

前記ネットワーク制御部が、複数のアドレスが設定されている 1 の前記コマン

ドをディスク制御部に送信し、

前記ディスク制御部が、前記コマンドを受信してこのコマンドに設定されている各アドレスに対応するデータ入出力をディスクドライブに対して行うようにすること、

を特徴とするディスク制御装置の制御方法。

【請求項 1 6】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、前記ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、前記回路基板に前記ネットワーク制御部および前記ディスク制御部が共通してアクセス可能なメモリを有するディスク制御装置を、

前記ネットワーク制御部および前記ディスク制御部が、設定されたタイミングで前記メモリに自身の動作状態を示す動作状態情報を更新し、

前記動作状態情報に基づいて前記ネットワーク制御部もしくは前記ディスク制御部に障害が発生したことを検知するようにすること、

を特徴とするディスク制御装置の制御方法。

【請求項 1 7】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、ディスク制御部はバックアップ装置との接続手段を備えるディスク制御装置を、

前記ネットワーク制御部が、前記ディスクドライブに記憶しているデータについてのバックアップ要求を外部装置から受信して前記ディスク制御部にバックアップコマンドを送信し、

前記ディスク制御部が、前記バックアップコマンドを受信すると前記ディスク



ドライブのデータについてのバックアップ指示をバックアップ装置に送信するようにすること、

を特徴とするディスク制御装置の制御方法。

【請求項 1 8】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、互いに通信可能に接続された複数の前記回路基板を備えるディスク制御装置を、

前記回路基板間でハートビートメッセージを交換することで 1 の前記回路基板に障害が生じたことを他の回路基板が検知し、

前記回路基板が他の前記回路基板についての障害を検知した場合に、障害となっている前記回路基板が行っている処理をその回路基板とは異なる他の回路基板に代行させるようにすること、

を特徴とするディスク制御装置の制御方法。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

この発明は、ディスク制御装置およびディスク制御装置の制御方法に関する。

【0 0 0 2】

【従来の技術】

SAN (Storage Area Network) を用いて構成されるストレージシステムの典型的な構成を図 1 2 に示す。ディスク制御装置 6 0 が提供する記憶領域にアクセスしようとする情報処理装置であるホストコンピュータ 2 0 は、LAN (Local Area Network) 5 0 を通じてサーバコンピュータ 7 0 にアクセスする。サーバコンピュータ 7 0 とディスク制御装置 6 0 とはファイバチャネル 8 0 により接続されている。

【0 0 0 3】

サーバコンピュータ 70 ではファイルシステム 75 が動作している。ホストコンピュータ 20 からサーバコンピュータ 70 へは、ファイル指定によるデータ入出力要求が送信される。ファイルシステム 75 はこのデータ入出力要求に基づいて S C S I 規格などに従った I / O コマンドを生成し、生成した I / O コマンドをファイバチャネル 80 を通じてディスク制御装置 60 に送信する。ディスク制御装置 60 はファイバチャネル 80 を通じて送られてくる I / O コマンドに従ってディスクドライブ 90 に対してデータ入出力を行う。ディスク制御装置 60 は、ディスクドライブ 90 から読み出したデータや処理完了報告などをサーバコンピュータ 70 に送信し、さらにサーバコンピュータ 70 からホストコンピュータ 20 に読み出したデータや処理完了報告が通知される。

【 0 0 0 4 】

【特許文献 1】

特開平 8 - 3 3 5 1 4 4 号公報

【 0 0 0 5 】

【発明が解決しようとする課題】

ところで、ファイバチャネル 80 による通信は、ファイバチャネル規格や S C S I 規格に従って行われるが、これらの規格は、時としてサーバコンピュータ 70 やディスク制御装置のハードウェアが有する潜在的な性能や能力を引き出すことの妨げとなる。またファイバチャネル 80 で通信を行うためには、ディスク制御装置 60 やサーバコンピュータ 70 内で稼働する C P U やメモリ間で行われる内部バスを通じた通信とファイバチャネル 80 を用いた通信との間でのプロトコル変換のための回路が必要であり、これにより装置の複雑化やコスト増を招く。またこの変換にかかる処理のオーバーヘッドはストレージシステムの性能を低下させる要因となる。

【 0 0 0 6 】

本発明は、このような事情に鑑みてなされたもので、ディスク制御装置およびディスク制御装置の制御方法を提供することを目的とする。

【 0 0 0 7 】

【課題を解決するための手段】

この目的を達成するための主たる発明にかかるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、

前記ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行う手段と、

前記ネットワーク制御部が複数のアドレスが設定されている 1 の前記コマンドをディスク制御部に送信する手段と、

前記ディスク制御部が前記コマンドを受信してこのコマンドに設定されている前記各アドレスに対応するデータ入出力をディスクドライブに対して行う手段と、

を備えることとする。

なお、本発明の他の特徴については、本明細書及び添付図面の記述により明らかにする。

【 0 0 0 8 】

【発明の実施の形態】

<開示の概要>

以下の開示により、少なくとももつぎのことが明らかにされる。

前記の発明において、ネットワーク制御部は、LANを通じてホストコンピュータ 2 0 から所定のネットワークプロトコルに従って送られてくるデータ入出力要求を受信するという、従来のストレージシステムにおけるサーバコンピュータが行っていた機能を提供する。またネットワーク制御部は、ウイルスチェックや SNMP (Simple Network Management Protocol)、クラスタ管理、ファイルに対するアクセス制限や時刻管理などのシステム管理の機能などを備えていることもある。また、ディスク制御部は、SANを通じてサーバコンピュータからコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うという、従来のストレージシステムにおけるディスク制御装置の機能を提供する。なお、データ入出力という場合には、ディスクドライブへのデータの書き

込みもしくはディスクドライブからの読み出しのうち少なくともいずれかが含まれる。前記コマンドは、例えば、後述する I / O コマンドが対応する。

【 0 0 0 9 】

このディスク制御装置では、ネットワーク制御部とディスク制御部とが同一回路基板上に形成され、両者がこの回路基板に設けられた P C I バス (Peripheral Component Interconnect Bus) などの内部バスで接続されていることで、ファイバチャネルによる通信の場合のようにプロトコルの制限に束縛されることなくネットワーク制御部とディスク制御部とは自由度の高い通信を行うことが可能である。

【 0 0 1 0 】

また、前記のようにネットワーク制御部が複数のアドレスが設定されている 1 の前記コマンドをディスク制御部に送信し、ディスク制御部が前記コマンドを受信してこのコマンドに設定されている各アドレスに対応するデータ入出力をディスクドライブに対して行うようにすることが可能である。

【 0 0 1 1 】

従って、従来のファイバチャネルによる通信では、あるデータ入出力要求を受信したサーバコンピュータがディスク制御装置に複数のコマンドを送信する必要があった処理を 1 のコマンドの送信により行うことができ、これにより通信にかかるオーバーヘッドが軽減されてストレージシステムの性能向上が図られる。

【 0 0 1 2 】

また、ネットワーク制御部とディスク制御部とが回路基板に設けられた内部バスを通じて接続することで、前述のようなファイバチャネルの通信と内部バスの通信との間の変換のための回路も必要なく、ストレージシステムの生産性の向上やコスト低減も図られる。

【 0 0 1 3 】

また、ディスク制御部がディスクドライブへのデータ入出力に関する処理について並行処理が可能な場合には、1 のコマンドで複数のアドレスに対応するデータ入出力に関する処理を並行して行うことが可能となり、これによりストレージシステムの性能向上が図られる。また、ネットワーク制御部とディスク制御部と

を同一回路基板に形成することで、これらが異なる回路基板で構成される場合に比べて製造工程の簡素化を図ることができる。

【 0 0 1 4 】

また、ネットワーク制御部は、ディスクドライブに入出力されるデータをファイル名により指定するデータ入出力要求を受け付ける仕組みを提供するファイルシステムが動作していることもあり、この場合、ネットワーク制御部はデータ入出力要求に設定されるファイル名に対応するデータのディスクドライブ上の記憶位置に対応するアドレスを生成し、このアドレスをコマンドに設定する。

【 0 0 1 5 】

本発明の一の態様であるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、前記回路基板に前記ネットワーク制御部および前記ディスク制御部が共通してアクセス可能なメモリを備え、前記ネットワーク制御部および前記ディスク制御部が設定されたタイミングで前記メモリに自身の動作状態を示す動作状態情報を更新する手段を備え、前記動作状態情報に基づいて前記ネットワーク制御部もしくは前記ディスク制御部に障害が発生したことを検知する手段を備えることとしている。

【 0 0 1 6 】

ここで、ネットワーク制御部が、ディスク制御部に送信したコマンドについての受信通知を取得できない場合に動作状態情報に基づいて送信先のディスク制御部の動作状態を調査し、その調査の結果に応じてそのコマンドを再びディスク制御部に送信するかどうかを判断するようにすることもできる。

【 0 0 1 7 】

また、ネットワーク制御部が、ディスク制御部に送信したコマンドについての受信通知を取得できない場合に動作状態情報に基づいて送信先の前記ディスク制御部の動作状態を調査し、ディスク制御装置が正常に動作していないと判断した

場合には、他のディスク制御部に前記コマンドを送信するようにすることもできる。また、障害を検知した場合にその旨を通知するユーザインタフェースを設けてもよい。さらに、障害を検知した場合に、障害が発生している前記ネットワーク制御部もしくは前記ネットワーク制御部に、再起動を要求する信号を送信するように構成してもよい。これによりフェールオーバを実現できる。

#### 【0018】

このようにネットワーク制御部とディスク制御部とを同一回路基板に形成し、さらに両者が共通してアクセス可能なメモリを設けることで、メモリにネットワーク制御部およびディスク制御部の動作状態情報を記憶しネットワーク制御部およびディスク制御部における障害を検知するようにするといった障害検知の仕組みを構成することができる。

#### 【0019】

そしてこの構成では、ネットワーク制御部やディスク制御部からの動作状態情報の更新、動作状態情報の参照、再起動信号の送信などの仕組みは、高速かつ信頼性の高い内部バスを用いた通信を介して行われる。このため、ストレージシステムの性能および信頼性を向上させることができる。

#### 【0020】

また、ネットワーク制御部が、前記コマンドをディスク制御部に送信するに際し前記コマンドの送信先となるディスク制御部の動作状態を動作状態情報から取得し、取得した動作状態に応じて前記コマンドをディスク制御部に送信するかどうかを決定するようにすることもできる。

#### 【0021】

本発明の一の態様であるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行い、ディスク制御部はバックアップ装置との接続手段を備え、ネットワーク制御部が、前記ディスクド

ライブに記憶しているデータについてのバックアップ要求を外部装置から受信して前記ディスク制御部にバックアップコマンドを送信し、前記ディスク制御部が、前記バックアップコマンドを受信すると前記ディスクドライブのデータについてのバックアップ指示をバックアップ装置に送信する手段を備えることとしている。

#### 【 0 0 2 2 】

従来、バックアップ装置はファイバチャネルを介してサーバコンピュータに接続する構成が一般的であり、この場合、サーバコンピュータはファイバチャネルを介してディスク制御装置からバックアップ装置にデータ転送を行う必要があった。しかしながら、このようにネットワーク制御部とディスク制御部とが内部バスにより接続する構成では、ネットワーク制御部からバックアップ指示コマンドをディスク制御部に送信するのみでディスク制御部によりバックアップを開始させることができる。そして、この構成ではディスク制御部からバックアップ装置に直接にバックアップ対象となるデータの転送が行われ、バックアップ中にネットワーク制御部にかかる負荷は大幅に低減されることになる。また、ファイバチャネル等の通信を介さない分、処理の高速化が図られる。

#### 【 0 0 2 3 】

本発明の一態様であるディスク制御装置は、ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを備え、ディスク制御部が、ネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うディスク制御装置において、互いに通信可能に接続された複数の前記回路基板を備え、前記回路基板間でハートビートメッセージを交換することで1の前記回路基板に障害が生じたことを他の回路基板が検知する手段を備え、前記回路基板が他の前記回路基板についての障害を検知した場合に、障害となっている前記回路基板が行っている処理をその回路基板とは異なる他の回路基板に代行させる手段を備えることとしている。

## 【 0 0 2 4 】

このように、ネットワーク制御部とディスク制御部とが形成された回路基板を複数有する構成では、これら回路基板の間でハートビートメッセージを交換し、他の回路基板に障害が生じた場合にその回路基板で行われていた処理を他の回路基板に代行させるようにすることで、フェールオーバーの仕組みを実現することができ、ストレージシステムの信頼性を向上させることができる。

## 【 0 0 2 5 】

## &lt;システム構成&gt;

図 1 に本発明の一実施例として説明するストレージシステムの構成を示す。

ディスク制御装置 1 0 に LAN 5 0 を介して一台以上のホストコンピュータ 2 0 が接続する。ホストコンピュータ 2 0 は、パーソナルコンピュータやワークステーション、汎用機などである。

## 【 0 0 2 6 】

ディスク制御装置 1 0 は、ネットワーク制御部 1 1 1、ディスク制御部 1 1 2、障害監視部 1 1 3、キャッシュメモリ 1 1 4、ディスクドライブ 1 1 5などを備えて構成される。なお、ディスクドライブ 1 1 5は、ディスク制御装置 1 0 が収容される筐体とは別筐体に収容されディスク制御装置 1 0 と通信手段を介して接続する構成であってもよい。

## 【 0 0 2 7 】

このディスク制御装置 1 0 では、ネットワーク制御部 1 1 1、ディスク制御部 1 1 2、障害監視部 1 1 3 が、同じ一枚の回路基板 1 1 7 上に形成されている。また、この回路基板 1 1 7 上には、さらに、LAN アダプタ (Lan Adaptor) もしくは NIC (Internet Interface Card) などに相当する機能を実現する回路であるネットワークインタフェース 1 1 8、DMA コントローラ (Direct Memory Access Controller) 1 1 9 が形成されている。

そして、回路基板 1 1 7 上に形成されたこれらの回路は、例えば、PCI バス (Peripheral Component Interconnect Bus) などの内部バス 3 0 により通信可能に結合されている。

## 【 0 0 2 8 】



ネットワーク制御部 1 1 1 は、CPU、メモリ（以下、「ネットワーク制御部用メモリ 1 2 1」と称する）などからなる。ネットワーク制御部 1 1 1 では、オペレーティングシステムが動作しており、このオペレーティングシステム上では、TCP/IP（登録商標）、NFS（Network File System）（登録商標）などのネットワークプロトコルに対応した通信を行うためのプログラムなど、各種のプログラムが動作する。

#### 【 0 0 2 9 】

ディスク制御部 1 1 2 は、CPU、メモリ（以下、「ディスク制御部用メモリ 1 2 2」と称する）、ディスクドライブ制御回路などからなる。また、ディスク制御部 1 1 2 は、ディスクドライブ 1 1 5 を RAID（Redundant Array of Inexpensive Disks）の方式で制御する機能を備えていることもある。

障害監視部 1 1 3 は、CPU、メモリ（以下、「障害監視部用メモリ 1 2 3」と称する）などからなる。

#### 【 0 0 3 0 】

また、ディスク制御装置 1 0 は、前記回路基板 1 1 7 上もしくはこれとは別の回路基板上に、バックアップ装置 1 8 4 が接続するためのインタフェース回路 1 8 3 を備える。なお、このインタフェース回路 1 8 3 は内部バス 3 0 に接続されている。インタフェース回路 1 8 3 に接続されるバックアップ装置 1 8 4 としては、DAT テープドライブ、DVD-RAM、MO、CD-R、ディスクドライブ、カセットテープなどがある。

#### 【 0 0 3 1 】

#### < データ入出力要求 >

ネットワーク制御部 1 1 1 は、ホストコンピュータ 2 0 から送信されてくるデータ入出力要求を LAN 5 0 を通じて受信すると、図 2 に示すようにこれをネットワーク制御部用メモリ 1 2 1 に確保されている受信バッファに記憶する。なお、図 2 ではホストコンピュータ 2 0 ごとに別の受信バッファを用意しているが、複数のホストコンピュータ 2 0 に共通の受信バッファを用意してもよい。

#### 【 0 0 3 2 】

ホストコンピュータ 2 0 から送られてくるデータ入出力要求には、それがデー

タ書き込み (Write) 要求であるのか、データ読み出し (Read) 要求であるのかなどを識別するコマンドコード、ファイル名、処理対象となるデータのファイル内の位置を特定するためのファイル内オフセットアドレスとデータサイズ、そのデータ入出力要求がデータ書き込み要求である場合に設定される書き込みデータ、などの情報が含まれている。

#### 【 0 0 3 3 】

ネットワーク制御部 1 1 1 は、受信バッファに記憶しているデータ入出力要求に基づいて I / O コマンドを生成し、生成した I / O コマンドを、内部バス 3 0 を通じてディスク制御部 1 1 2 に送信する。ディスク制御部 1 1 2 は、ネットワーク制御部 1 1 1 から I / O コマンドを受信すると、この I / O コマンドに対応するデータ入出力をディスクドライブ 1 1 5 に対して行う。

#### 【 0 0 3 4 】

#### < ファイル管理情報 >

ネットワーク制御部 1 1 1 からディスク制御部 1 1 2 に送信される I / O コマンドは、データ入出力要求と、ファイル管理情報に基づいて生成される。ファイル管理情報はディスクドライブ 1 1 5 に記憶されている。

#### 【 0 0 3 5 】

図 3 にファイル管理情報の一例を示す。ファイル管理情報には、ファイル名に対応させて、そのファイルのデータサイズであるファイルサイズ、当該ファイル名に対応するデータが複数の記憶領域に分割されて記憶されている場合における分割数を示すデータ領域数、各連続領域のディスクドライブ 1 1 5 上の記憶位置を特定するデータアドレスやデータサイズなどが設定されている。なお、図 4 に示すようにファイル管理情報は、処理性能の向上のため、ネットワーク制御部用メモリ 1 2 1 に記憶 ( キャッシング ) されることもある。

#### 【 0 0 3 6 】

#### < I / O コマンド >

図 5 は、データ入出力要求に基づいて I / O コマンドが生成される仕組みを説明している。ネットワーク制御部 1 1 1 は、まずデータ入出力要求のファイル内オフセットアドレスをファイル管理情報に対照し、そのデータ入出力要求が処理

対象とするファイルに関するデータのディスクドライブ 1 1 5 上の記憶位置に対応するデータアドレスとデータサイズとを求める。前述したように、1 のファイルに対応するデータは、ディスクドライブ 1 1 5 上の連続する記憶領域に纏まって記憶されている場合もあるし、複数の記憶領域に分割されて記憶されている場合もある。図 5 は、1 のファイルに対応するデータが 2 つの記憶領域に分割されて記憶されている場合である。

#### 【 0 0 3 7 】

ここでデータ入出力要求の処理対象であるファイルが、ディスクドライブ 1 1 5 上の連続する記憶領域に纏まって記憶されている場合、ネットワーク制御部 1 1 1 はそのデータ入出力要求に対応する I / O コマンドとして、その記憶領域を指定するための、1 の論理アドレス ( L B A ( Logical Block Address ) ) とブロック数とが設定された I / O コマンドを生成する。ここで論理アドレスとは、ディスクドライブの記憶領域に区画設定された論理的な記憶領域 ( 以下、「論理ユニット」もしくは「 L U ( Logical Unit ) 」と称する) におけるデータの記憶位置を指定する論理的なアドレスである。

#### 【 0 0 3 8 】

一方、データ入出力要求の処理対象であるファイルが、ディスクドライブ 1 1 5 の複数の記憶領域に分割されて記憶されている場合、ネットワーク制御部 1 1 1 は、分割された各記憶領域のそれぞれを指定する複数の論理アドレスとブロック数との組合せを設定した I / O コマンドを生成する。またこの場合には、ファイルがいくつの記憶領域に分割されているかを示す分割数が I / O コマンドのリスト数の欄に設定される。

#### 【 0 0 3 9 】

図 6 は、生成された I / O コマンドがネットワーク制御部用メモリ 1 2 1 に記憶されている様子を示している。なお、この図では I / O コマンドは L U ごとにネットワーク制御部用メモリ 1 2 1 上に用意されたコマンドテーブルに管理されているが、これに限られるわけではない。

#### 【 0 0 4 0 】

< I / O コマンドの生成 >

つぎにネットワーク制御部 1 1 1 がデータ入出力要求に基づいて I / O コマンドを生成する処理について、図 7 に示すフローチャートとともに説明する。

ネットワーク制御部 1 1 1 は、受信バッファにデータ入出力要求が存在する場合 (S711)、受信バッファからデータ入出力要求を一つ取り出し (S712)、データ入出力要求に設定されているファイル名に対応するファイルをネットワーク制御部用メモリ 1 2 1 のファイル管理情報から検索する (S713)。

#### 【 0 0 4 1 】

ここで対応するファイルがネットワーク制御部用メモリ 1 2 1 のファイル管理情報に存在する場合 (S714: YES) には、(S716) に進む。他方、対応するファイルが存在しない場合 (S714: NO) はファイル管理情報をディスクドライブ 1 1 5 から読み出す (S715)。

#### 【 0 0 4 2 】

(S716) ではデータ入出力要求のオフセットアドレスとデータサイズに対応するディスクドライブ 1 1 5 のデータアドレスとデータサイズとを求める。ここで前述したようにそのファイルを構成するデータがディスクドライブ 1 1 5 上の複数の記憶領域に分割されて記憶されている場合には、それぞれの領域に対応するデータアドレスとデータサイズとを求める。そして算出したデータアドレスとデータサイズとが設定された I / O コマンドをネットワーク制御部用メモリ 1 2 1 上に生成する (S717)。

#### 【 0 0 4 3 】

I / O コマンドがネットワーク制御部用メモリ 1 2 1 に生成される様子を図 8 に示している。図 8 (a) は、複数のホストコンピュータ 2 0 から 1 の論理ボリュームについての I / O コマンドが生成される場合における I / O コマンドの生成の様子である。図 8 (b) は 1 のホストコンピュータ 2 0 が複数の論理ボリュームにアクセスする場合における I / O コマンドの生成の様子である。

#### 【 0 0 4 4 】

図 7 の (S718) では、ネットワーク制御部 1 1 1 は、つぎに処理すべきデータ入出力要求が存在するかどうかを受信バッファから調べる。他のホストコンピュータ 2 0 の受信バッファが存在する場合には、(S712) からの処理が繰り返され

る。

一方、つぎに処理すべきデータ入出力要求が受信バッファに存在しない場合（S578:NO）、ネットワーク制御部 1 1 1 は、I/O コマンドの実行要求を適宜なタイミングでディスク制御部 1 1 2 に送信する（S719）。

なお、（S719）の処理は（S718）の処理を待たずに行うように構成することもできる。また以上の処理は、受信バッファが複数存在する場合には各受信バッファについて行われる（S719）。

#### 【 0 0 4 5 】

#### < I / O コマンドの実行 >

つぎに、ディスク制御部 1 1 2 による I / O コマンドの実行について、図 9 に示すフローチャートとともに説明する。

ディスク制御部 1 1 2 は、ネットワーク制御部 1 1 1 から受信した I / O コマンドの実行要求を、ディスク制御部用メモリ 1 2 2 に記憶している。ディスク制御部 1 1 2 は、ディスク制御部用メモリ 1 2 2 に I / O コマンドの実行要求が存在するかどうかを監視している（S911）。ディスク制御部 1 1 2 は、ディスク制御部用メモリ 1 2 2 に I / O コマンドの実行要求が存在する場合、ディスク制御部用メモリ 1 2 2 から I / O コマンドを 1 つ読み出し（S912）、そのコマンドに設定されているコマンドコードから、そのコマンドがデータ書き込みのコマンドであるのか、データ読み出しのコマンドであるのかを調べる（S913）。ここで読み出しコマンドの場合には、（S914）に進む。一方、データ書き込みコマンドの場合には（S931）に進む。

#### 【 0 0 4 6 】

（S914）において n には初期値として 1 が設定されているものとする。ディスク制御部 1 1 2 は、L B A 番号 n の L B A およびデータサイズを読み出す（S915）。つぎにディスク制御部 1 1 2 は、その L B A およびデータサイズに対応するデータがキャッシュメモリ 1 1 4 に存在するかどうかを調べ（S916）、存在する場合（S916:YES）にはキャッシュメモリ 1 1 4 からそのデータを読み出してネットワーク用メモリに転送する（S917）。一方、キャッシュメモリ 1 1 4 に存在しない場合、ディスク制御部 1 1 2 はディスクドライブ 1 1 5 に読み出し要求を送

信する (S918)。つぎにディスク制御部 1 1 2 は L B A 番号 n をインクリメントし (S919)、インクリメント後の n の値を I / O コマンドのリスト数と比較する (S920)。ここで n がリスト数以下の場合 (S920: YES) には (S915) に進む。一方、n がリスト数よりも大きい場合 (S920: NO) には (S921) に進む。

## 【 0 0 4 7 】

(S921) では、(S918) において読み出し要求を送信した場合 (S921: YES)、ディスクドライブ 1 1 5 がその要求に対応するデータを読み出すのを待ってから (S923) に進む (S922)。一方、(S918) において読み出し要求を送信していない場合には (S923) に進む。

## 【 0 0 4 8 】

(S923) では、ディスクドライブ 1 1 5 からデータが読み出されると、そのデータをキャッシュメモリ 1 1 4 に記憶するとともにネットワーク制御部用メモリ 1 2 1 へ転送し、ディスク制御部 1 1 2 はネットワーク制御部 1 1 1 にデータ入出力要求についての処理を完了した旨のステータスを報告する。

## 【 0 0 4 9 】

一方、(S913) において I / O コマンドが書き込み要求であった場合には、(S931) に進む。(S931) において、n には初期値として 1 が設定されているものとする。(S932) において、ディスク制御部 1 1 2 は L B A 番号 n の L B A およびデータサイズを読み出す。

## 【 0 0 5 0 】

つぎにディスク制御部 1 1 2 は、その L B A およびデータサイズに対応するデータがキャッシュメモリ 1 1 4 に存在するかどうかを調べ (S933)、存在しない場合 (S933: NO) には、キャッシュメモリ 1 1 4 に記憶領域を確保し (S934)、その記憶領域に書き込みデータを書き込む (S735)。

一方、存在する場合 (S933: YES) には、そのデータに書き込みデータを上書きする (S936)。

(S937) では、その書き込みデータに対応させてキャッシュメモリに記憶されている当該書き込みデータについてのデステージフラグをオンにする。なお、デステージフラグがオンになっている書き込みデータは、適宜なタイミングで、キ

キャッシュメモリ 1 1 4 からディスクドライブ 1 1 5 にデステージされる。

【 0 0 5 1 】

つぎにディスク制御部 1 1 2 は L B A 番号  $n$  をインクリメントし (S938)、インクリメント後の  $n$  が I / O コマンドのリスト数に設定されている値と比較する (S939)。ここで  $n$  がリスト数以下の場合 (S939: YES) には (S932) に進む。また、 $n$  がリスト数よりも大きい場合 (S939: NO) には (S923) に進み、ディスク制御部はネットワーク制御部にデータ入出力要求についての処理を完了した旨のステータスを報告する (S940)。

【 0 0 5 2 】

以上のように、ネットワーク制御部 1 1 1 からディスク制御部 1 1 2 に複数のアドレスが設定されている 1 のコマンドを送信するだけで、ディスク制御部 1 1 2 が複数のアドレスに対応するデータ入出力をディスクドライブ 1 1 5 に対して行う。このためコマンドの処理にかかるオーバーヘッドが削減されてストレージシステムの性能向上が図られる。

【 0 0 5 3 】

<バックアップ処理>

つぎにホストコンピュータ 2 0 から送信されてくるバックアップ要求に応じてディスク制御装置 1 0 が実行するデータバックアップ処理について説明する。

ネットワーク制御部 1 1 1 で動作するオペレーティングシステム上では、ホストコンピュータ 2 0 から送信されてくるバックアップ要求を受信した場合にディスクドライブ 1 1 5 に記憶しているデータを、図 1 に示すバックアップ装置 1 8 4 のバックアップメディアに記憶する処理を実行する、バックアッププログラムが動作している。

【 0 0 5 4 】

バックアップ要求には、ファイル名や論理ユニット名などの情報により、バックアップ対象となるデータが指定されている。バックアップ要求を受信すると、バックアッププログラムはバックアップ要求に指定されている情報に基づいて、バックアップ対象となるデータのディスクドライブ 1 1 5 上の記憶位置を特定するアドレス及びデータサイズを求め、これらを設定したバックアップ指示コマン

ドを内部バス 3 0 を通じてディスク制御部 1 1 2 に送信する。

【 0 0 5 5 】

ディスク制御部 1 1 2 は、バックアップ指示コマンドを受信すると、このコマンドに指定されているアドレスとデータサイズとで指定される記憶領域に格納されているデータと、バックアップの開始を指示するバックアップ開始コマンドとをバックアップ装置 1 8 4 に送信する。バックアップ開始コマンドを受信したバックアップ装置 1 8 4 は、このコマンドと共に送られてくるバックアップ対象のデータをバックアップメディアに記録する。

【 0 0 5 6 】

ディスク制御部 1 1 2 は、バックアップ対象データのバックアップメディアへの記録が終了した旨の通知をバックアップ装置 1 8 4 から受信すると、内部バス 3 0 を通じてバックアッププログラムに完了通知を送信する。バックアッププログラムは完了通知を受信すると、ホストコンピュータ 2 0 にバックアップ完了報告を送信する。

【 0 0 5 7 】

ところで、従来、バックアップ装置 1 8 4 はファイバチャネルを介してサーバコンピュータに接続する構成が一般的であり、この場合、サーバコンピュータはファイバチャネルを介してディスク制御装置からバックアップ装置にデータ転送を行う必要があった。

【 0 0 5 8 】

しかしながら、本実施例のようにネットワーク制御部 1 1 1 とディスク制御部 1 1 2 とが内部バス 3 0 により接続する構成では、ネットワーク制御部 1 1 1 からバックアップ指示コマンドをディスク制御部 1 1 2 に送信するのみでディスク制御部 1 1 2 によりバックアップを開始させることができる。そして、この構成ではディスク制御部 1 1 2 からバックアップ装置 1 8 4 に直接的にバックアップ対象となるデータの転送が行われ、バックアップ中にネットワーク制御部 1 1 1 にかかる負荷は大幅に軽減されることになる。また、ファイバチャネル等の通信を介さない分、処理の高速化も図られる。

【 0 0 5 9 】



### ＜複数の回路基板＞

フェールオーバさせること等を目的として、ディスク制御装置 1 0 に同一の機能を備える複数の回路基板が実装されることが行われているが、ネットワーク制御部 1 1 1 とディスク制御部 1 1 2 とが形成された 1 の回路基板を、1 台のディスク制御装置 1 0 に複数枚実装することもできる。図 1 0 は、そのように複数枚の前記回路基板が実装されたディスク制御装置 1 0 を用いて構成したストレージシステムである。

#### 【0 0 6 0】

図 1 0 において、各回路基板 1 1 7 上には、ネットワークインタフェース 1 1 8、ネットワーク制御部 1 1 1、ディスク制御部 1 1 2 および DMA コントローラ 1 1 9 が形成されており、これらは P C I バスなどの内部バス 3 0 に接続している。各回路基板 1 1 7 の内部バス 3 0 は互いに接続されている。

#### 【0 0 6 1】

また、ディスク制御装置 1 0 には、各回路基板 1 1 7 とは別に、障害監視部 1 1 3 および共用メモリ 1 4 1 が形成された回路基板 1 2 7 が実装されている。なお、この回路基板 1 2 7 の障害監視部 1 1 3 および共用メモリ 1 4 1 も、接続ライン 3 1 を介して内部バス 3 0 に接続している。

#### 【0 0 6 2】

ここでこのような構成からなるディスク制御装置 1 0 では、各回路基板 1 1 7 に形成されているネットワーク制御部 1 1 1 間で、内部バス 3 0 を通じてハートビートメッセージを交換して互いに動作状態を監視することで、フェールオーバの仕組みを実現することができる。

#### 【0 0 6 3】

以下、この仕組みについて、図 1 1 ～図 1 3 に示すフローチャートとともに説明する。

まず、前提として、共用メモリ 1 4 1 には、各回路基板のネットワーク制御部 1 1 1 の動作状態に関するステータス情報が記憶されているものとする。なお、動作に関するステータス情報とは、例えば、ネットワーク制御部 1 1 1 が、正常に動作しているかどうかを示すステータス情報である。

## 【 0 0 6 4 】

このステータス情報は、定期的もしくは不定期などの設定されたタイミングで、各回路基板 1 1 7 のネットワーク制御部 1 1 1 により内部バス 3 0 を通じて書き込まれる。なお、このステータス情報は、障害監視部 1 1 3 が内部バス 3 0 を通じて各ネットワーク制御部 1 1 1 に動作状態を問い合わせ、その応答として取得した各ネットワーク制御部 1 1 1 の動作状態を、障害監視部 1 1 3 が内部バス 3 0 を通じて間接的に共用メモリ 1 4 1 に書き込むこともある。

## 【 0 0 6 5 】

図 1 1 は、ある回路基板 1 1 7 のネットワーク制御部 1 1 1 が、他の回路基板 1 1 7 の他のネットワーク制御部 1 1 1 に対し、ハートビートメッセージを要求する側として動作する場合における当該ネットワーク制御部 1 1 1 の処理を説明するフローチャートである。

## 【 0 0 6 6 】

ネットワーク制御部 1 1 1 は、内部バス 3 0 を通じて共用メモリ 1 4 1 にアクセスしてステータス情報を調べ、正常に動作しているネットワーク制御部 1 1 1 を探索する (S1111)。そして、正常に動作している他のネットワーク制御部 1 1 1 が探索されると、そのネットワーク制御部 1 1 1 に対し、内部バス 3 0 を通じてハートビートメッセージを要求するメッセージを送信し (S1112)、そのネットワーク制御部 1 1 1 からハートビートメッセージが送られてくるのを一定時間待つ (S1113)。

## 【 0 0 6 7 】

ここで他のネットワーク制御部 1 1 1 からハートビートメッセージが送られてきた場合 (S1114: YES) には、リトライカウンタに「0」を設定し (S1115)、一定時間待機した後 (S1116)、再び (S1111) からの処理に戻る。

## 【 0 0 6 8 】

一方、(S1114) において、他のネットワーク制御部 1 1 1 からハートビートメッセージが受信できなかった場合 (S1114: NO) には、リトライカウンタに「1」を加算し (S1117)、リトライカウンタをあらかじめ設定されているリトライオーバー閾値と比較する (S1118)。

## 【 0 0 6 9 】

ここでリトライカウンタの値がリトライオーバ閾値を超えている場合（S1118：YES）には、内部バス 3 0 を通じて障害監視部 1 1 3 に、他のネットワーク制御部 1 1 1 に障害が発生している旨のメッセージを送信する（S1119）。

なお、このメッセージには、その障害がハードウェアの障害に起因するのか、ソフトウェアに起因するのかを示す情報や、あるネットワーク制御部 1 1 1 からのハートビートメッセージが中断している旨を通知する情報など、障害の内容に関する情報も付帯される。ネットワーク制御部 1 1 1 は、そのメッセージに対して障害監視部 1 1 3 から応答があると（S1120）、他のネットワーク制御部 1 1 1 の処理を代行する、フェールオーバ処理を開始する（S1121）。

## 【 0 0 7 0 】

一方、（S1118）の処理において、リトライカウンタの値がリトライオーバ閾値を超えていない場合（S1118：NO）には、リトライカウンタに「0」を設定し（S1115）、再び（S1111）からの処理に戻る。

## 【 0 0 7 1 】

図 1 2 は、ある回路基板 1 1 7 のネットワーク制御部 1 1 1 が、ハートビートメッセージを受信したり、障害監視部 1 1 3 からのメッセージを受信した場合における処理を説明するフローチャートである。

ネットワーク制御部 1 1 1 は、内部バス 3 0 を通じて送られてくるメッセージを受信すると（S1211）、そのメッセージがハートビートメッセージであるかどうかを調べる（S1212）。

ここで受信したメッセージがハートビートメッセージである場合（S1212：YES）、ネットワーク制御部 1 1 1 は、そのハートビートメッセージを送信したネットワーク制御部 1 1 1 に対し、ハートビートメッセージを送信する（S1213）。

## 【 0 0 7 2 】

一方、受信したメッセージがハートビートメッセージ以外のメッセージであった場合（S1212：YES）は、ネットワーク制御部 1 1 1 はそのメッセージが障害監視部 1 1 3 から送られてくる、他のネットワーク制御部 1 1 1 が動作していない旨を通知するメッセージであるかどうかを調べる（S1214）。

(S1214)において、そのメッセージが、他のネットワーク制御部 1 1 1 が動作していない旨を通知する障害監視部 1 1 3 からのメッセージであった場合には (S1214: YES)、そのネットワーク制御部 1 1 1 との間の通信を抑止して (S1215)、必要な場合にはそのネットワーク制御部 1 1 1 に関するフェールオーバー処理を開始する (S1216)。

## 【 0 0 7 3 】

図 1 3 は、障害制御部 1 3 の処理を説明するフローチャートである。障害制御部 1 3 は、図 1 1 における (S1119) の通知、すなわち、あるネットワーク制御部 1 1 1 に障害が発生している旨のメッセージを受信すると (S1311: YES)、そのメッセージに付帯する情報に基づいて、その障害通知がハードウェアの障害によるものであるのかを判断する (S1312)。

## 【 0 0 7 4 】

(S1312)において、ハードウェアの障害であると判断した場合には (S1312: YES)、障害監視部 1 1 3 は、前記メッセージにより特定されるネットワーク制御部 1 1 1 が動作していない旨のステータス情報を共用メモリ 1 4 1 に書き込む (S1313)。そして、障害監視部 1 1 3 は、前記メッセージで特定されるネットワーク制御部 1 1 1 に障害が発生している旨を記載したメッセージを、障害が発生していないネットワーク制御部 1 1 1 に送信する (S1314)。

## 【 0 0 7 5 】

一方、(S1312)において、ハードウェアの障害でないと判断した場合には (S1312: NO)、障害監視部 1 1 3 は、受信したメッセージが、ネットワーク制御部 1 1 1 からハートビートメッセージが送られてきていない旨を通知するメッセージであるかどうかを調べる (S1315)。そして、障害監視部 1 1 3 は、受信したメッセージが、ネットワーク制御部 1 1 1 からハートビートメッセージが送られてきていない旨を通知するメッセージであった場合には、内部バス 3 0 を通じてそのネットワーク制御部 1 1 1 を制御してそのネットワーク制御部 1 1 1 の動作を停止させた (S1316) 後、(S1313)からの処理に進む。

## 【 0 0 7 6 】

以上のようにしてフェールオーバーの仕組みを実現することができる。ちなみに

、LANを介してハートビートメッセージを通信する場合には、ハートビートメッセージの通信によりLANに負荷がかかり、また、この負荷を回避しようとするれば、ハートビートメッセージのための専用のLANを設ける必要があるが、内部バス30によりハートビートメッセージを通信する仕組みとした場合にはこのような問題や煩わしさが一切生じないという利点がある。

【0077】

#### ＜障害監視＞

処理能力の向上や可用性の向上などを目的として、ディスク制御装置10に複数のネットワーク制御部111やディスク制御部112が実装されることがある。

このような場合、複数のネットワーク制御部111やディスク制御部112を、図14に示すように同一の回路基板140上に形成することで、処理能力の向上や信頼性の向上を図ることができる。なお、回路基板140には、ネットワーク制御部111およびディスク制御部112が共通してアクセス可能な共用メモリ141が形成されている。

【0078】

共用メモリ141には、ネットワーク制御部111やディスク制御部112が内部バス30を通じて書き込む時刻情報を記憶するための記憶領域（以下、「タイマテーブル」と称する）が確保される。タイマテーブルには、ネットワーク制御部111やディスク制御部112の動作状態に関する情報が設定される。ここでは動作状態情報としてその書き込みが行われた時刻が書き込まれる。また、ネットワーク制御部111やディスク制御部112からタイマテーブルへの書き込みは、障害監視に必要とされる設定されたタイミング（例えば1秒間隔）で行われる。

【0079】

図15は、ネットワーク制御部111がディスク制御部112にI/Oコマンドを送信するに際し、ネットワーク制御部111により行われる障害検知の仕組みを説明するフローチャートである。

ネットワーク制御部111は、ディスク制御部112にI/Oコマンドを送信

するにあたり、事前にタイマテーブルを参照してその I / O コマンドの送信先のディスク制御部 1 1 2 が正常に動作しているかどうかを調査する (S1511, S1512)。なお、この調査は、例えば、タイマの更新が直前の所定秒数以内に行われていたかどうかを調べることにより行われる。

#### 【 0 0 8 0 】

この調査により I / O コマンドを送信しようとするディスク制御部 1 1 2 に対応するタイマテーブルの更新が行われていない (正常に動作していない) 場合には (S1512: NO)、ネットワーク制御部 1 1 1 はその処理を中断してホストコンピュータ 2 0 にエラーを報告し、また正常に動作しているディスク制御部 1 1 2 を検索し (S1513)、検索された他のディスク制御部 1 1 2 に I / O コマンドの送信を試みる (S1514)。

一方、(S1512) において I / O コマンドを送信しようとするディスク制御部 1 1 2 についてタイマテーブルの更新が行われている (正常に動作している) 場合には (S1512: YES)、そのディスク制御部 1 1 2 にコマンドを送信する (S1515)。

#### 【 0 0 8 1 】

ここで前記調査によりネットワーク制御部 1 1 1 が I / O コマンドの送信先のディスク制御部 1 1 2 が正常に動作していると判断 (S1512: YES) し、I / O コマンドをそのディスク制御部 1 1 2 に送信した (S1514, S1515) にもかかわらず、その I / O コマンドが送信先に受け付けられず、タイムアウトとなってしまうことがあるが、I / O コマンドを送信した後にディスク制御部 1 1 2 に障害が発生した可能性もあるため、この場合は (S1511) に進む。

一方、(S1516) において、タイムアウトにならずにネットワーク制御部 1 1 1 がディスク制御部 1 1 2 から I / O コマンドの受領通知を受信すれば処理が終了する (S1517)。

#### 【 0 0 8 2 】

以上の障害監視の仕組みによれば、ネットワーク制御部 1 1 1 からディスク制御部 1 1 2 に対する I / O コマンドを確実にディスク制御部 1 1 2 に引き渡すことが可能となり、これによりストレージシステムの信頼性を向上させることがで

きる。

【 0 0 8 3 】

< 障害監視部 >

ところで、前出の図 1 4 に示す回路基板 1 4 0 には、各ネットワーク制御部 1 1 1 および各ディスク制御部 1 1 2 に対応するタイマテーブルの時刻が更新されているかどうかを監視する障害監視部 1 3 が形成されている。障害監視部 1 3 は、例えば、一定時間以上、時刻の更新を行っていないネットワーク制御部 1 1 1 もしくはディスク制御部 1 1 2 を検知し、S N M P に従って外部装置やユーザインタフェースに障害の発生を通知する。

【 0 0 8 4 】

また障害監視部 1 3 は、時刻が更新されていない原因がネットワーク制御部 1 1 1 やディスク制御部 1 1 2 のファームウェアの障害によるものであると判断した場合、ネットワーク制御部 1 1 1 やディスク制御部 1 1 2 に内部バス 3 0 を通じてリセット信号（例えば、Power on reset 信号）を送信する。また、障害監視部 1 3 は、内部バス 3 0 を通じて送られてくるネットワーク制御部 1 1 1 やディスク制御部 1 1 2 からの通知によっても、ネットワーク制御部 1 1 1 やディスク制御部 1 1 2 に障害の発生を認知する。

【 0 0 8 5 】

以上のように、ネットワーク制御部 1 1 1、ディスク制御部 1 1 2、共用メモリ 1 4 1、障害監視部 1 3 が、それぞれ同じ回路基板上に形成されていることで、ネットワーク制御部 1 1 1 やディスク制御部 1 1 2 からの共用メモリ 1 4 1 のタイマテーブルへの時刻の書き込み、ネットワーク制御部 1 1 1 によるタイマテーブルの参照、障害監視部 1 3 によるタイマテーブルの時刻の更新状態の監視、ネットワーク制御部 1 1 1 やディスク制御部 1 1 2 へのリセット信号の送信などの処理は、全て内部バス 3 0 を通じて行われることになる。従って、これらの処理は高速かつ高信頼に行われ、ディスク制御装置 1 0 の処理の高速化や信頼性の向上が図られる。

【 0 0 8 6 】

以上、本発明に係るディスク制御装置等について説明してきたが、上記した発

明の実施の形態は、本発明の理解を容易にするためのものであり、本発明を限定するものではない。本発明は、その趣旨を逸脱することなく、変更、改良され得ると共に、本発明にはその等価物が含まれることは勿論である。

【0087】

【発明の効果】

本発明によれば、ディスク制御装置およびディスク制御装置の制御方法を提供することができる。

【図面の簡単な説明】

【図1】 本発明の一実施例による、ストレージシステムの構成を示す図である。

【図2】 本発明の一実施例による、データ入出力要求がネットワークデータ用メモリの受信バッファに記憶される様子を説明する図である。

【図3】 本発明の一実施例による、ファイル管理情報を示す図である。

【図4】 本発明の一実施例による、ファイル管理情報が制御用メモリに記憶される様子を説明する図である。

【図5】 本発明の一実施例による、データ入出力要求に基づいてI/Oコマンドが生成される仕組みを説明する図である。

【図6】 本発明の一実施例による、生成されたI/Oコマンドが制御用メモリに記憶されている様子を説明する図である。

【図7】 本発明の一実施例による、ネットワーク制御部がデータ入出力要求に基づいてI/Oコマンドを生成する処理を説明するフローチャートである。

【図8】 (a)、(b)は、それぞれ本発明の一実施例による、I/Oコマンドが制御用メモリに生成される様子を説明する図である。

【図9】 本発明の一実施例による、I/Oコマンドの実行を説明するフローチャートを示す図である。

【図10】 本発明の一実施例による、複数枚の回路基板が実装されたディスク制御装置を用いて構成されるストレージシステムの構成を示す図である。

【図11】 本発明の一実施例による、ある回路基板のネットワーク制御部が、他の回路基板の他のネットワーク制御部に対し、ハートビートメッセージを



要求する側として動作する場合における当該ネットワーク制御部の処理を説明するフローチャートを示す図である。

【図 1 2】 本発明の一実施例による、ある回路基板のネットワーク制御部が、ハートビートメッセージを受信したり、障害監視部からのメッセージを受信した場合における処理を説明するフローチャートを示す図である。

【図 1 3】 本発明の一実施例による、障害制御部の処理を説明するフローチャートである。

【図 1 4】 本発明の一実施例による、複数のネットワーク制御部およびディスク制御部が形成された回路基板を示す図である。

【図 1 5】 本発明の一実施例による、障害検知の仕組みを説明するフローチャートを示す図である。

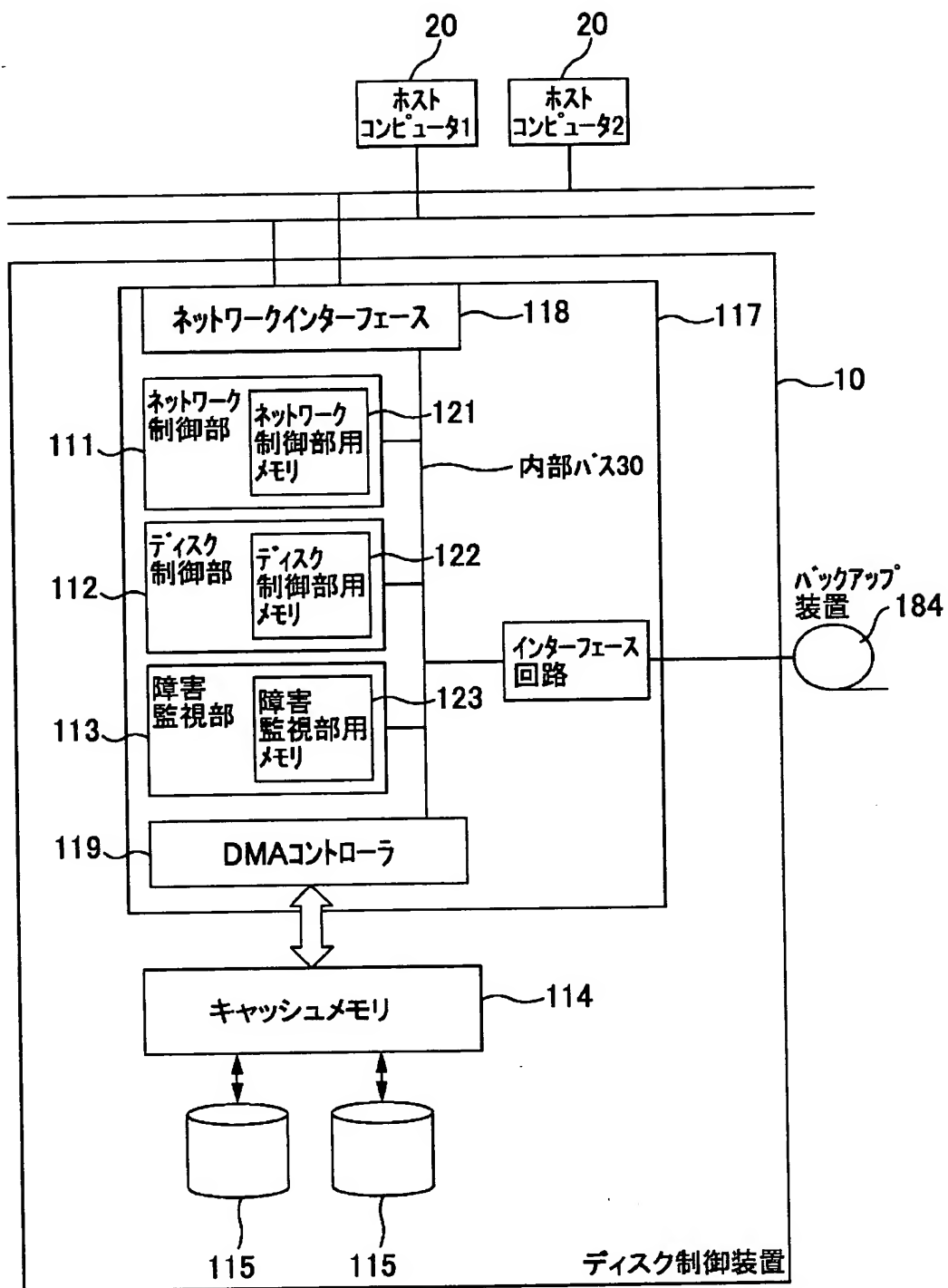
【図 1 6】 従来の典型的なストレージシステムの構成を示す図である。

【符号の説明】

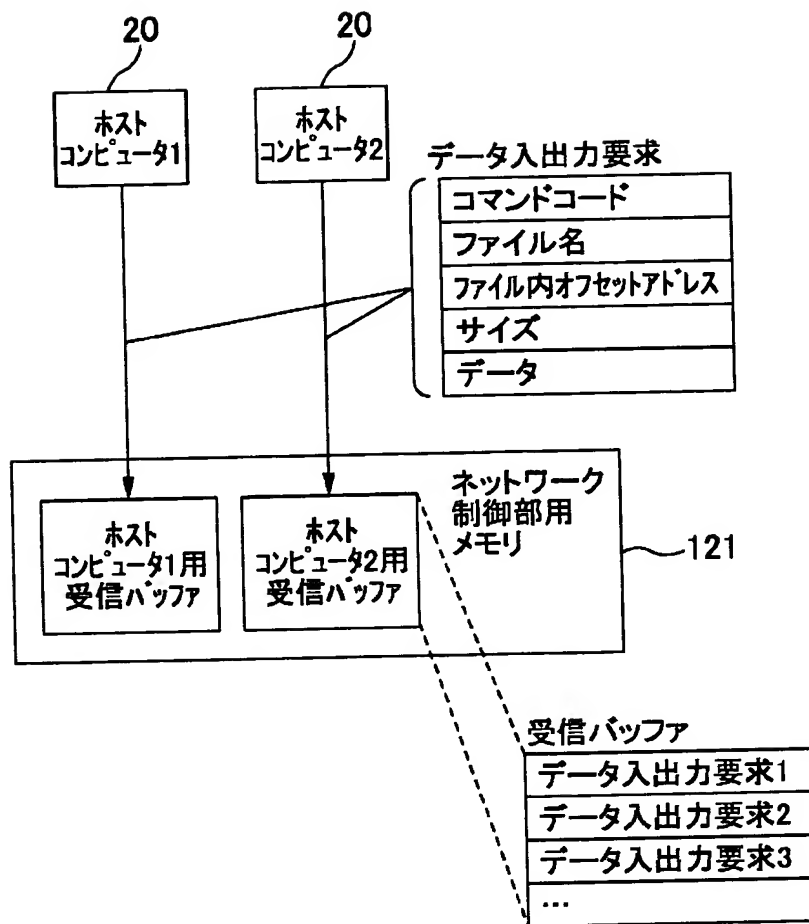
- 1 0 ディスク制御装置
- 1 8 回路基板
- 2 0 ホストコンピュータ
- 3 0 内部バス
- 5 0 LAN (ネットワーク)
- 1 1 1 ネットワーク制御部
- 1 1 2 ディスク制御部
- 1 1 3 障害監視部
- 1 1 4 キャッシュメモリ
- 1 1 5 ディスクドライブ
- 1 2 1 ネットワーク制御部用メモリ
- 1 2 2 ディスク制御部用メモリ
- 1 4 0 回路基板
- 1 4 1 共用メモリ
- 1 8 4 バックアップ装置

【書類名】 図面

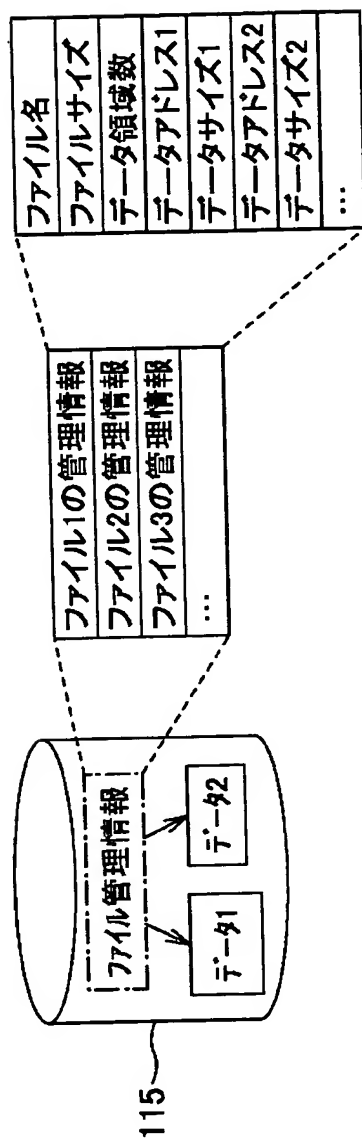
【図 1】



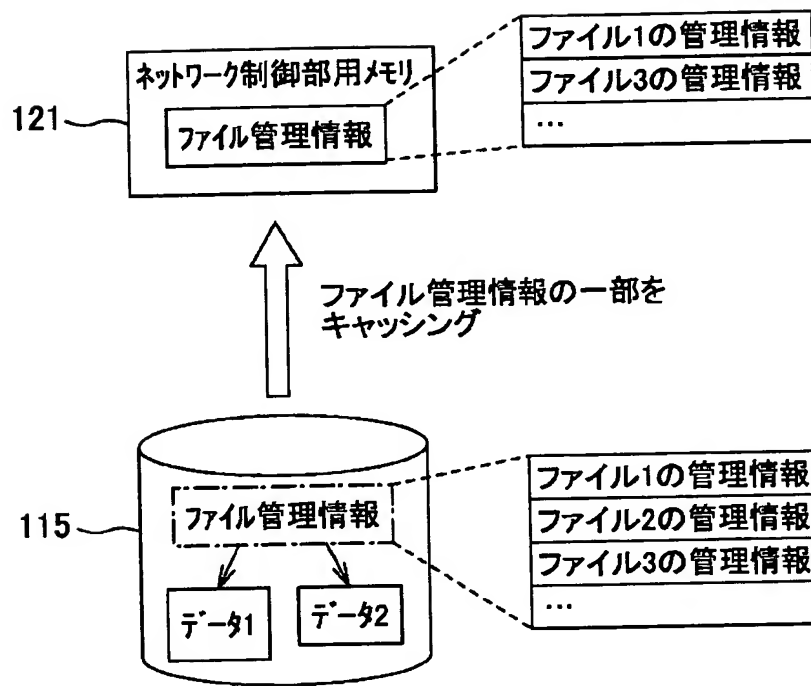
【図 2】



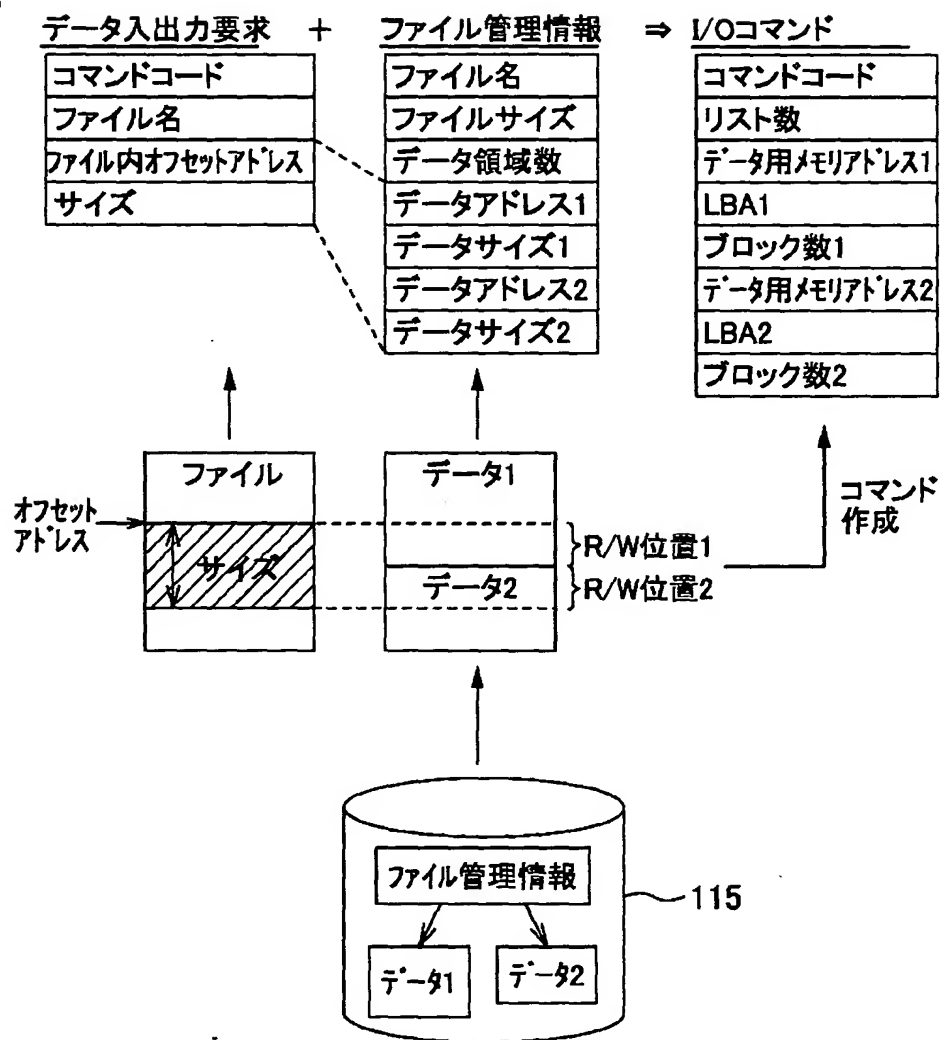
【図 3】



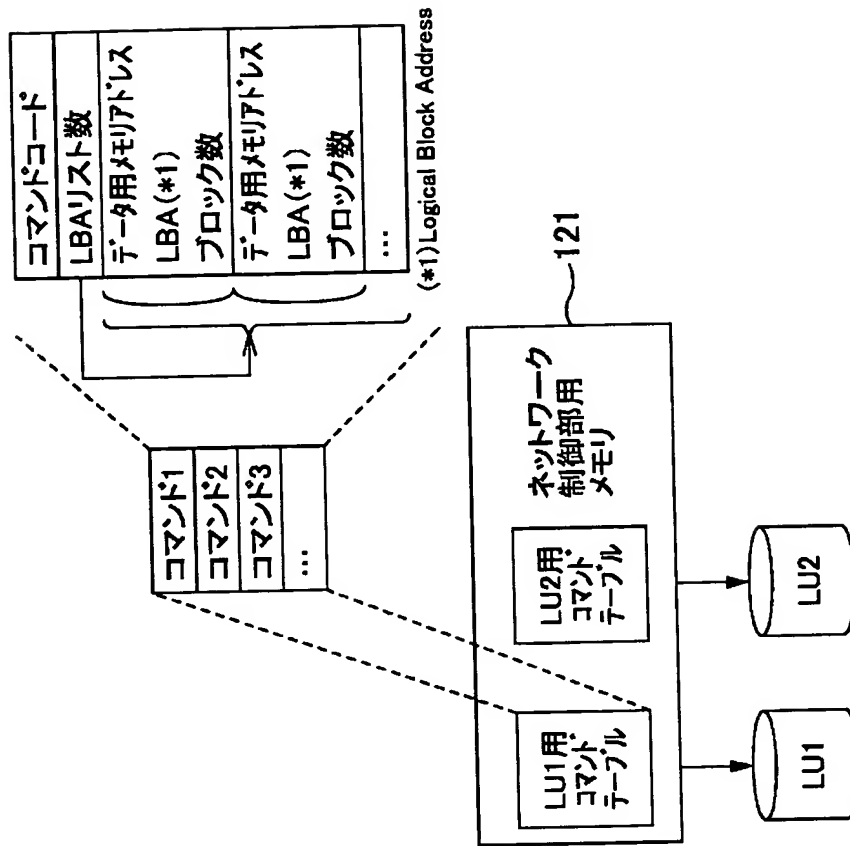
【図 4】



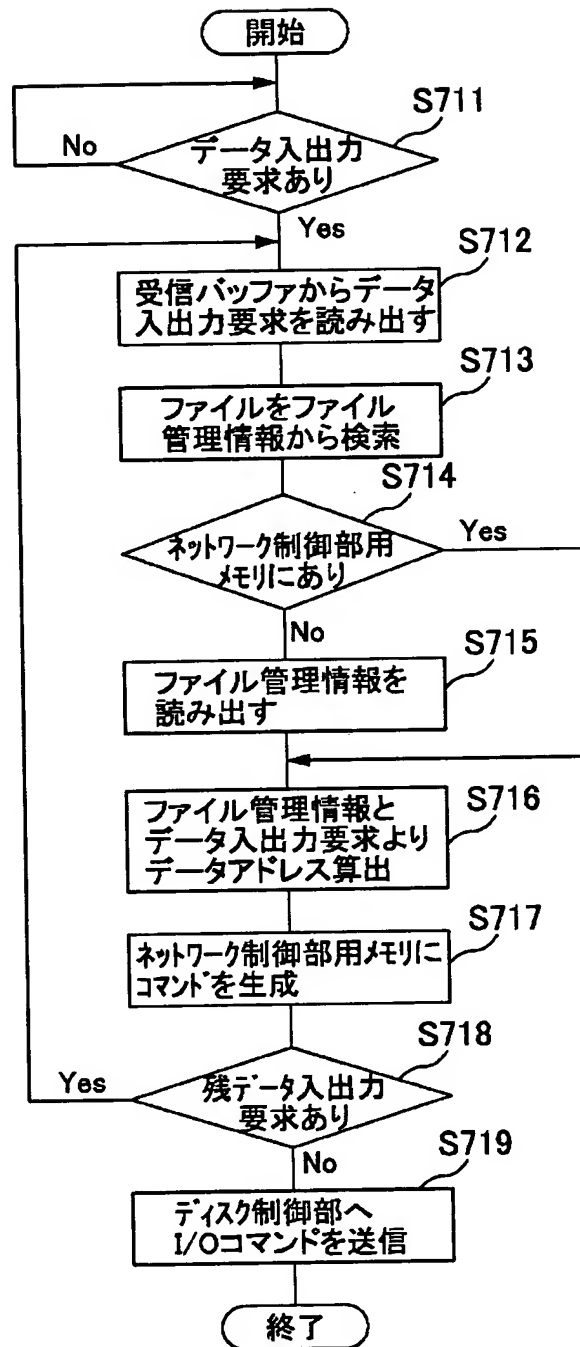
【図 5】



【図 6】

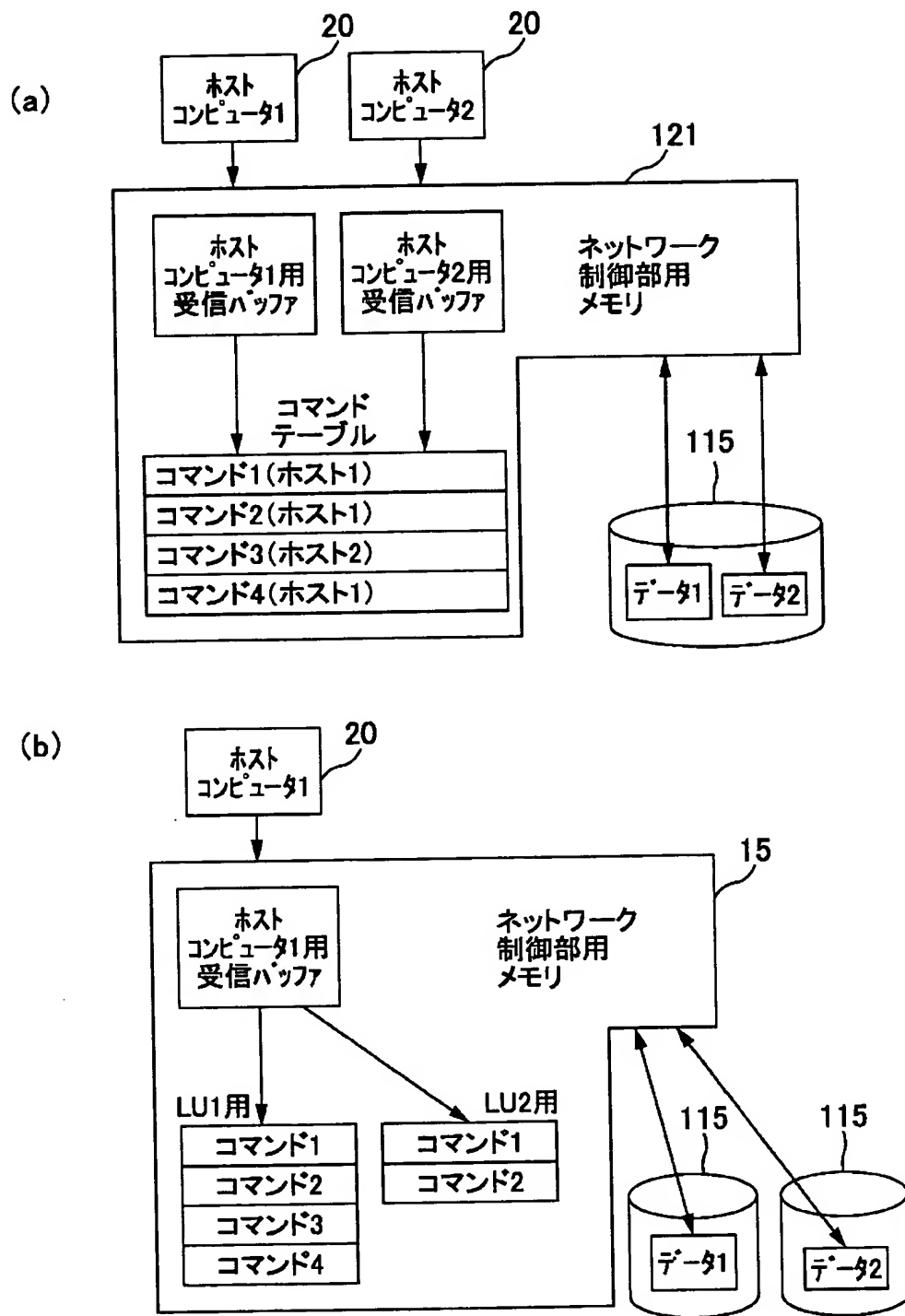


【図 7】

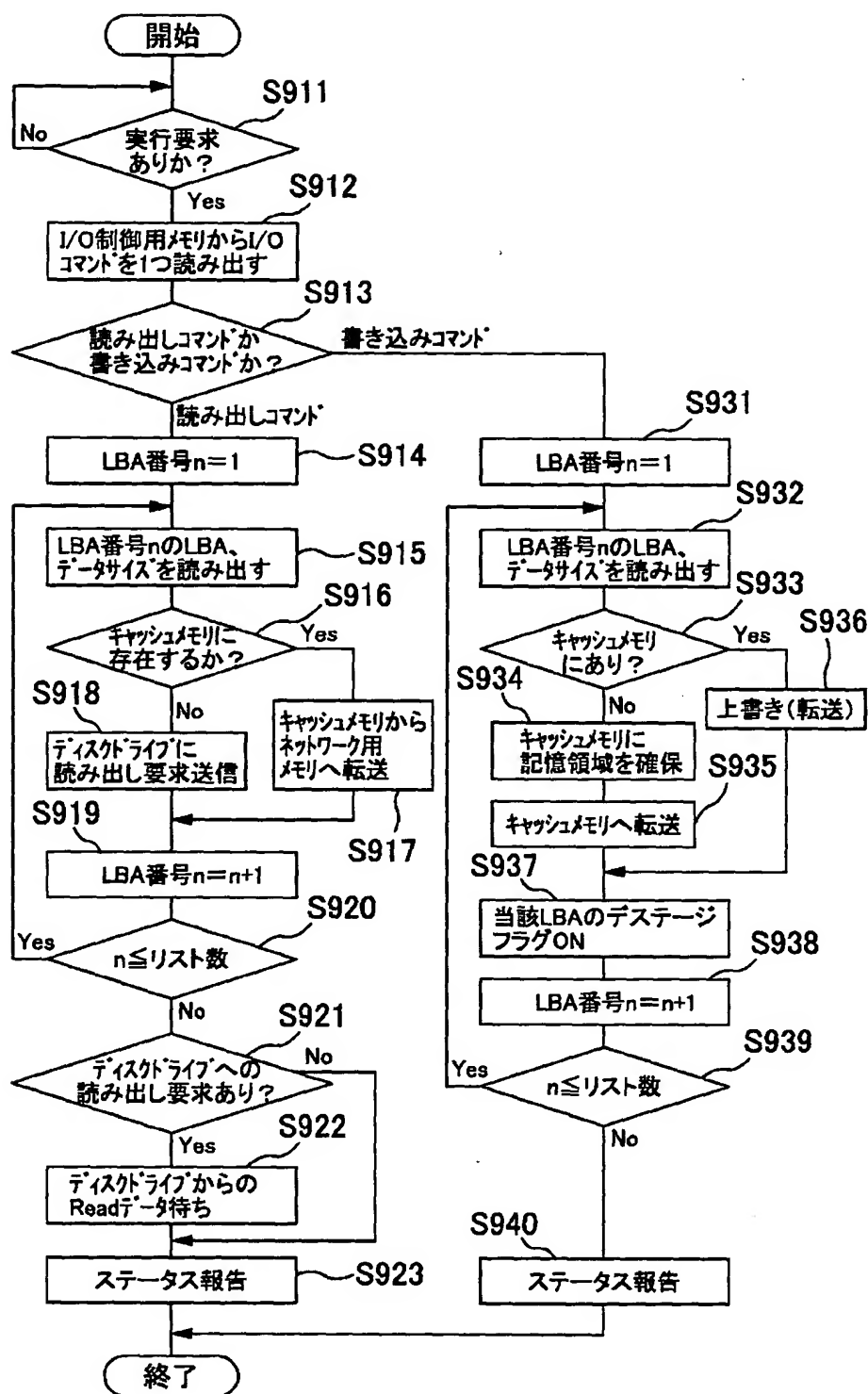




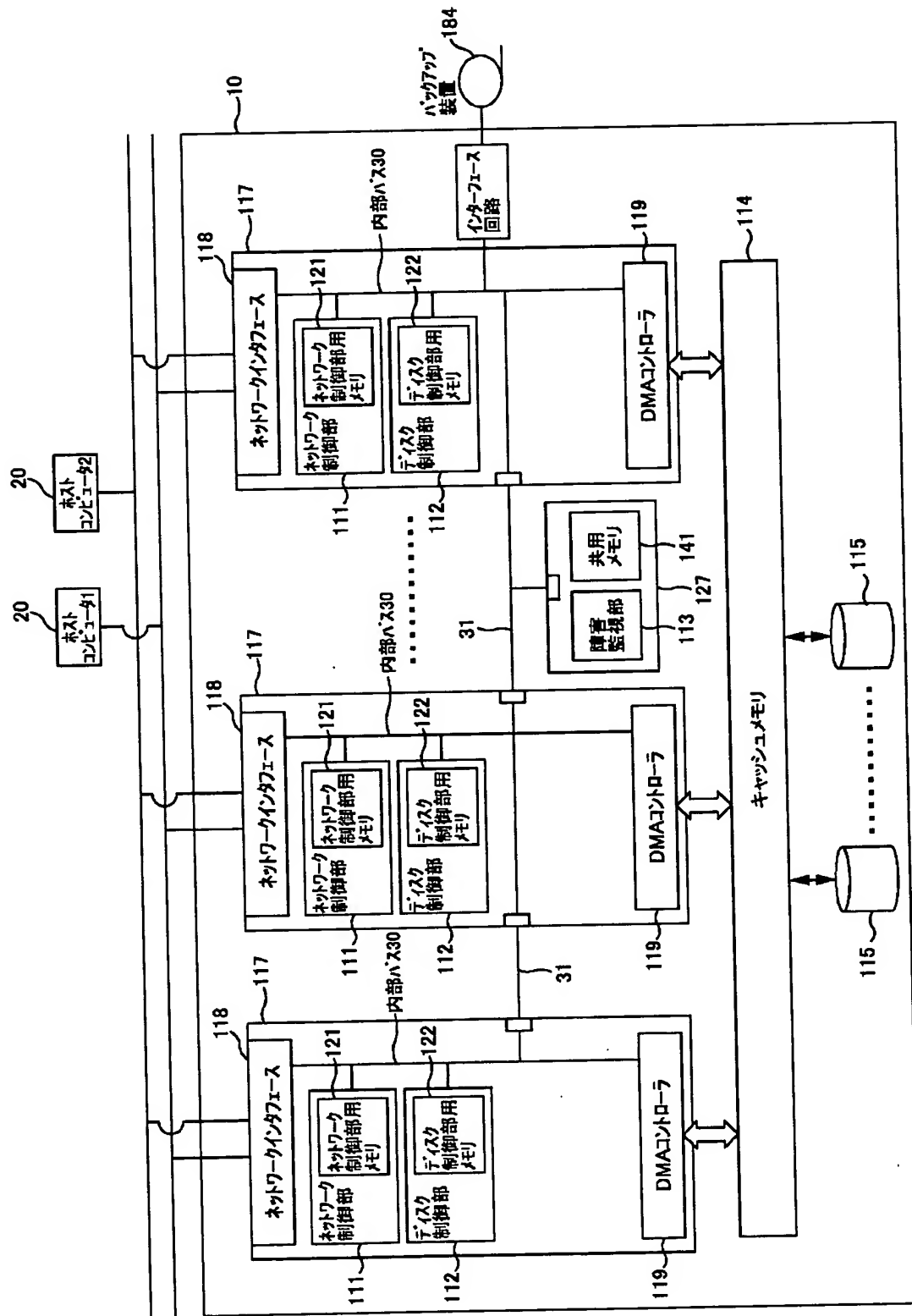
【図 8】



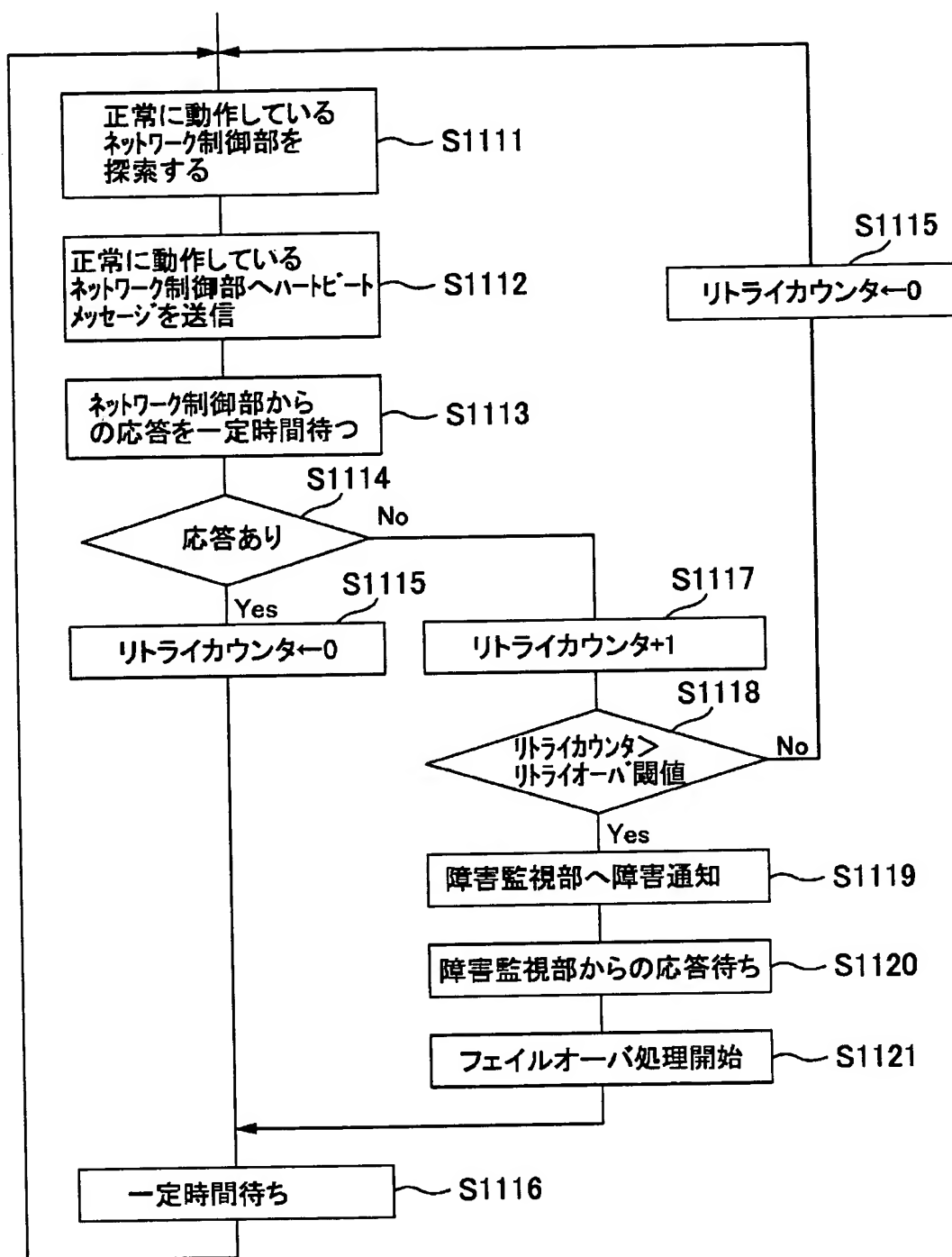
【図9】



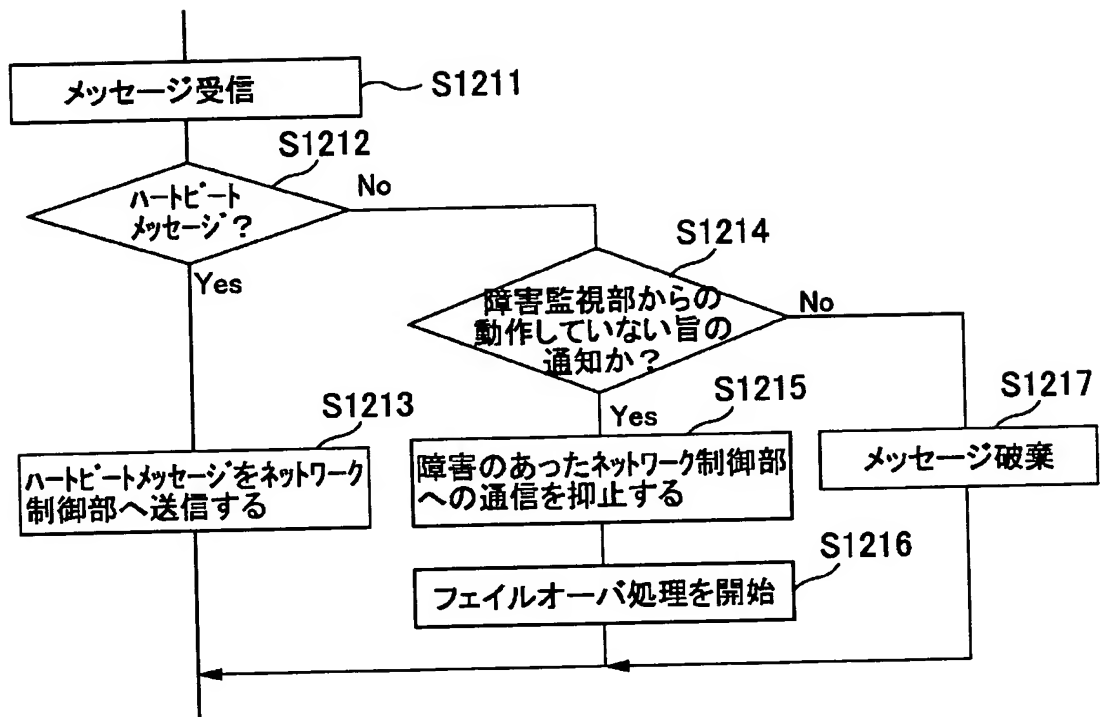
【図10】



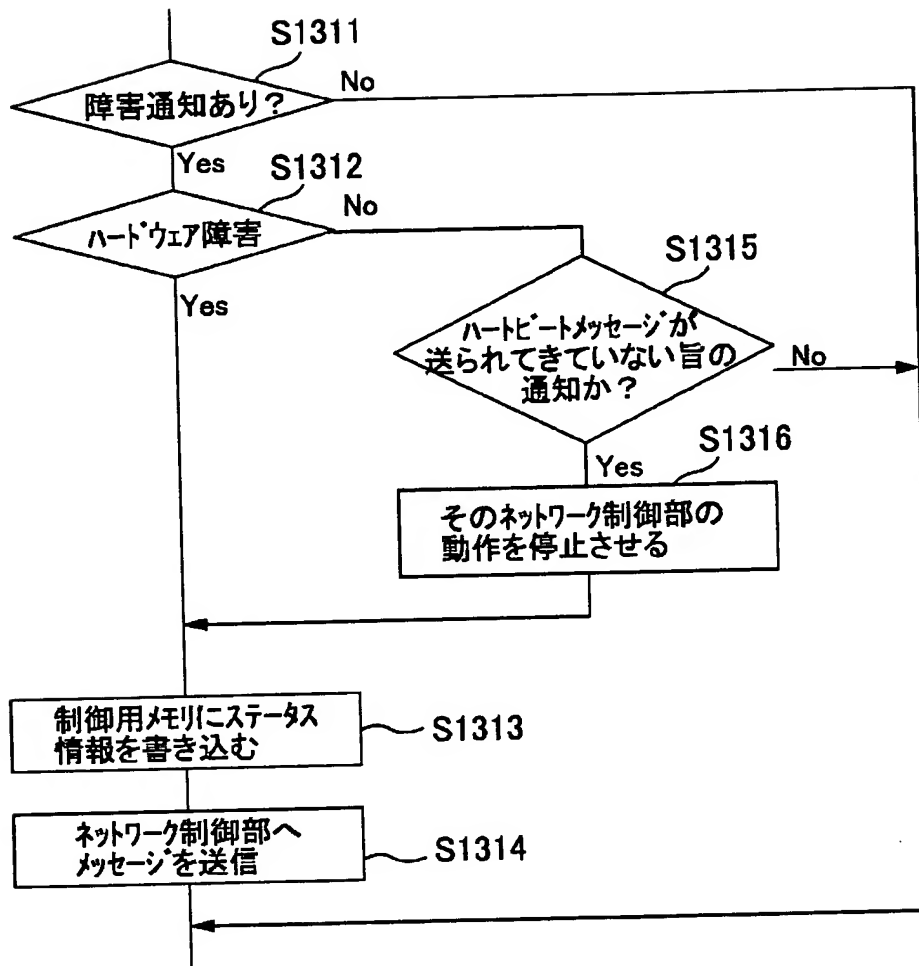
【図 1 1】



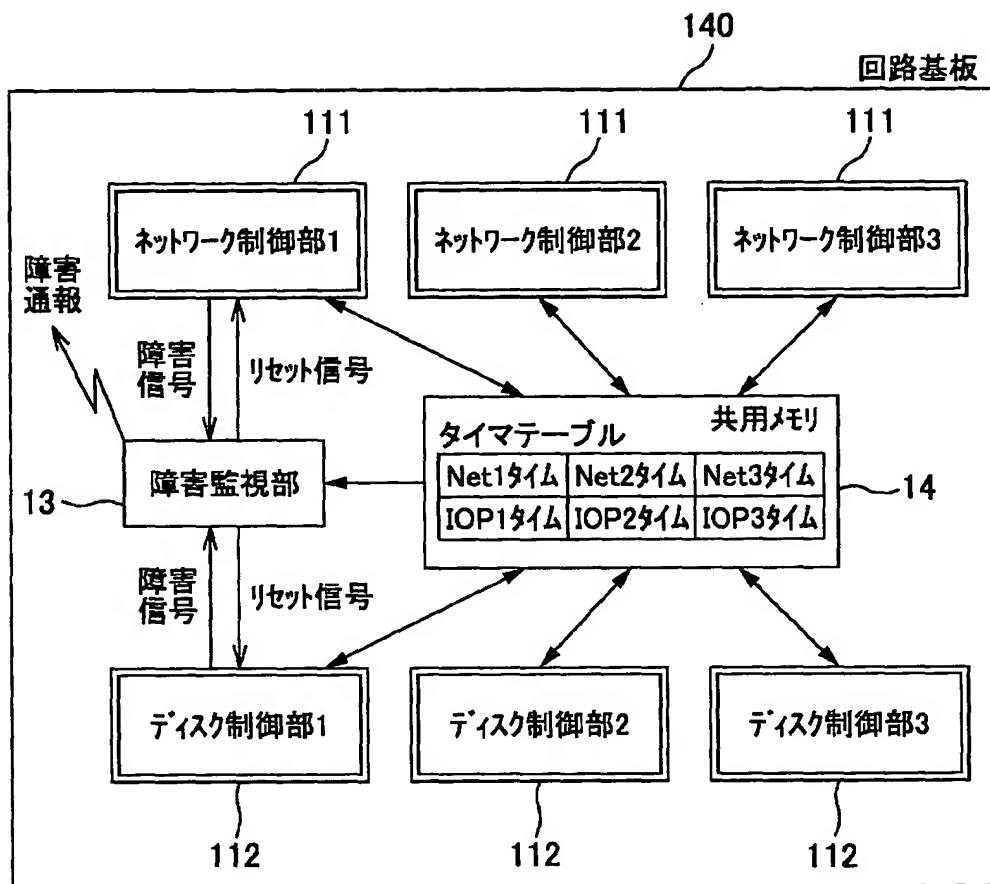
【図12】



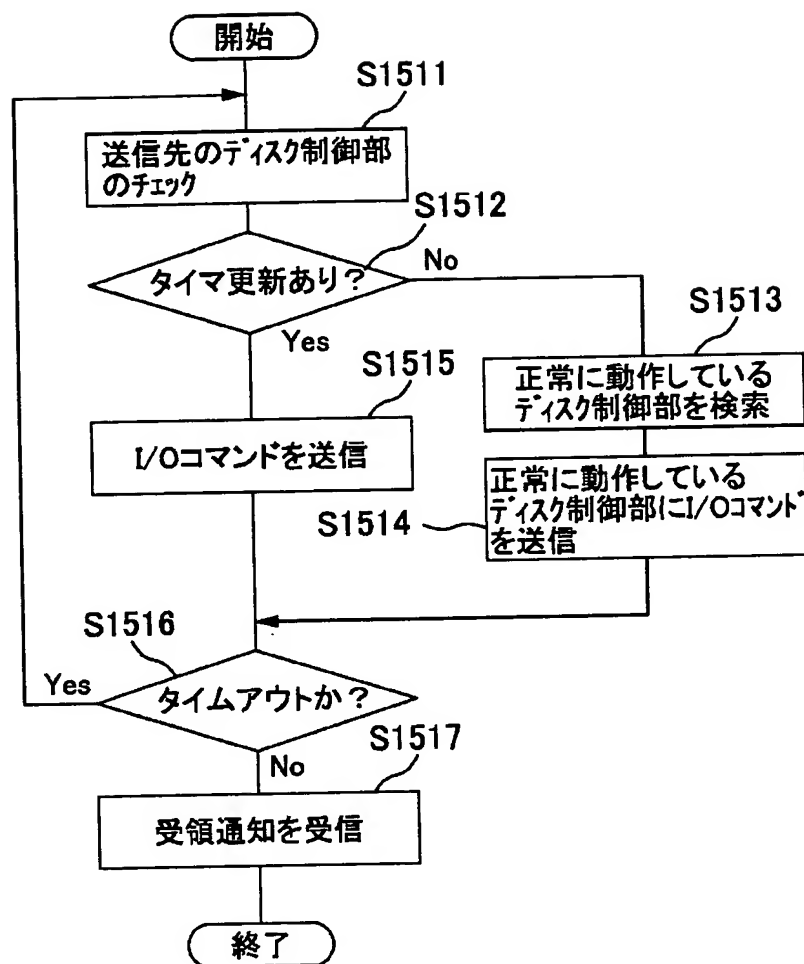
【図 1 3】



【図14】

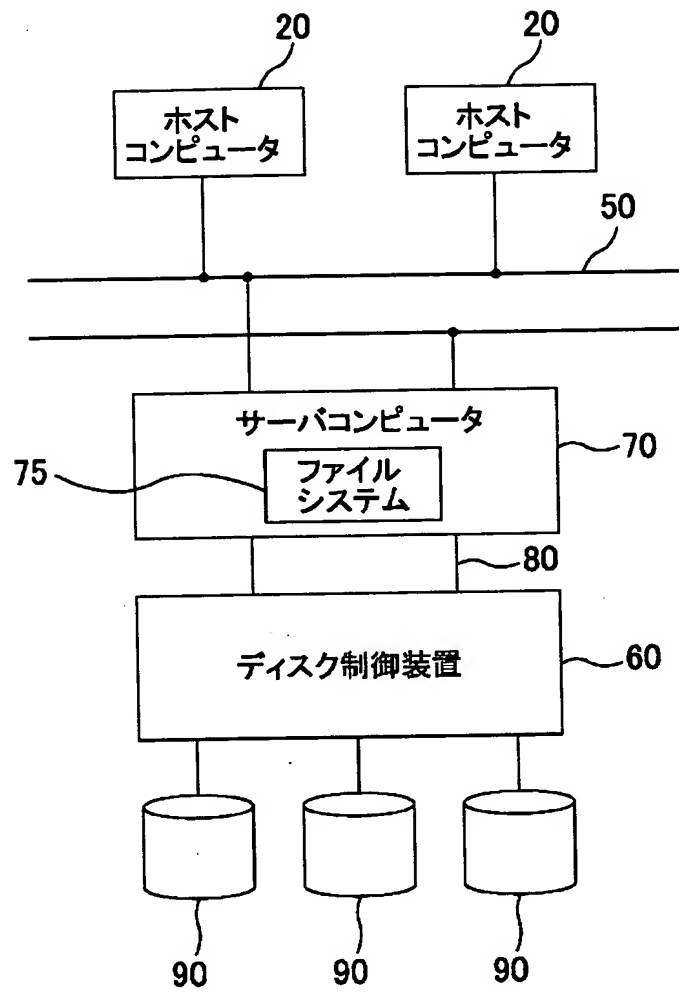


【図 1 5】





【図 1 6】



【書類名】 要約書

【要約】

【解決手段】 ネットワークを通じて外部装置から送られてくるデータ入出力要求を受信するネットワーク制御部と、ネットワーク制御部と同一回路基板に形成され当該基板に設けられた内部バスによりネットワーク制御部と接続するディスク制御部とを有し、ディスク制御部がネットワーク制御部から内部バスを通じて送信されてくるコマンドを受信してこのコマンドに応じてディスクドライブに対するデータ入出力を行うように構成したディスク制御装置を提供する。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000005108]

1. 変更年月日 1990年 8月31日

[変更理由] 新規登録

住 所 東京都千代田区神田駿河台4丁目6番地

氏 名 株式会社日立製作所